Common words between two random strings

Philippe Jacquet

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Common words between two random strings

Philippe Jacquet

N° 5899

Avril 2006

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--5899--FR+ENG

Common words between two random strings

Philippe Jacquet*

Thème COM — Systèmes communicants
Projets Hipercom

Rapport de recherche n° 5899 — Avril 2006 — 11 pages

Abstract: We investigate the problem of enumerating the words that are common to two random strings. We show that when the source models are memoriless that the number of common words is sublinear in the length of the sequence, and linear when the source models are exactly the same. We draw the same conclusions for the number of common nodes in the associated respective suffix trees.

Key-words: words, strings, information theory, suffix tree, complex analysis

This is a note

This is a second note

* Projet Hipercom, INRIA Rocquencourt

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Mots communs entre deux séquences aléatoires

Résumé : Nous regardons le problème de l'énumération des mots communs entre deux séquences aléatoires. Lorsque les sources sont sans mémoire, nous montrons que le nombre de mots communs est sous linéaire par rapport à la longueur des séquences et est linéaires quand les modèles de source sont exactement les mêmes. Nous tirons les mêmes conclusions pour le nombre de noeuds communs entre les arbres suffix associés.

Mots-clés : mots, séquences, théorie de l'information, arbre suffix, analyse complexe

1 Introduction

We plan to work on the average number of common words between two random sequences of length n. We investigate the model of memoriless sources (for instance binary, but easily extendable to larger alphabet). The Markov source model can also be extended the same way but this will not be investigated in the current paper. In [1], Janson, Lonardi and Szpankowski have investigated the complexity of a random string, i.e. the average number of distinct words in a random sequence. In [2], Jacquet and Szpankowski have investigated the number of occurence of a given word in a random string, result later extended by Régnier and Szpankowski on Markov model [3, 4]. The analysis is based on the convergence between the suffix tree statistics to the independent tries statistics.

In the present paper we analyze the average number of common words in two random sequences, generated from two different random sources. We will show that the average number of common words between two sequences of length n is in $\Theta(n)$ when the sequence are generated from the same source model and is in $\Theta(n^\kappa)$ for some $\kappa < 1$ when the source models are different. We make use of analytic information theory and will use [2] as inspiration source for the general results used in this paper. We also introduce the two-variable Mellin transform for extracting asympotic expansions. We also investigate the average number of common nodes between the suffix trees of two random sequences as a powerful tool to analyse the analogy between two models.

2 Autocorrelations in sequence

In [2] we have shown that the distribution of the number $H_n(\sigma)$ of occurence of a word σ in a sequence of length n is $P_n(v)$ satisfies the identity:

$$\sum_k P(H_n(\sigma) = k)v^k = [z^n]\left(1 - z - \frac{(v-1)p(\sigma)z^{|\sigma|}}{1-(v-1)a_\sigma(z)}\right)^{-1} \tag{1}$$

where $p(\sigma)$ is the probability of the word σ in the considered source model, $a_\sigma(z)$ is the autocorrelation polynomial of the word σ in the considered source model.

The important fact is that we prove that excepted for a set S of words σ we have $\sum_k P(H_n(\sigma) = k)v^k = P_n^T(v,\sigma)(1 + O(n^{-\epsilon}))$ for some $\epsilon > 0$, with

$$P_n^T(v,\sigma) = ((1-(v-1)p(\sigma))^n \tag{2}$$

And for all n, $\sum_{\sigma\in S,|\sigma|} p(\sigma) \leq O(\rho^n)$ for some $\rho < 1$.

We notice that $P_n^T(v,\sigma)$ is the p.g.f. of the number of sequences among n independent independent sequences that accept σ as prefix.

3 Warmup: sequence complexity

The complexity C_n of a sequence [1] is the number of distinc words in a sequence of length n. The number of words is $\frac{(n+1)n}{2}$ but some of them are identical, for example there are n empty words. Therefore we have

$$C_n = \frac{(n+1)n}{2} - \sum_{\sigma} \sum_{k \geq 2} (k-1) P(N_n(\sigma) = k) \tag{3}$$

Or we have

$$\begin{aligned} C_n &= \frac{(n+1)n}{2} - \sum_{\sigma} ((P_n^T)'(1,\sigma) - (P_n^T)'(0,\sigma)) + \\ &\quad \sum_{\sigma} P_n^T(1,\sigma) - P_n^T(0,\sigma) - (P_n^T)'(0,\sigma) + O(n^{1-\epsilon}) \end{aligned}$$

We identify in $\sum_\sigma P_n^T(1,\sigma) - P_n^T(0,\sigma) - (P_n^T)'(0,\sigma)$ the number of internal nodes in a tries made of n independent sequence, and in $\sum_\sigma ((P_n^T)'(1,\sigma) - (P_n^T)'(0,\sigma))$ the external path length of the trie. The former is in $\frac{n}{h}$ and the latter is in $\frac{n \log n}{h}$ with h being the entropy rate of the random source, with possible fluctuating terms $Q(\log n)$ with periodic function $Q(.)$ when the logarithms of the symbol probabilities are integer proportional. We got back the result of [1]:

$$C_n = \frac{(n+1)n}{2} - \frac{(\log n - 1)n}{h} + nQ(\log n) + o(n) \tag{4}$$

4 Common words analysis

Our aim is to give the average number K_n of common words between two random sequences of same length n but not on same source model. We have

$$K_n = \sum_{\sigma} P(N_n^1(\sigma) \geq 1) P(N_n^2(\sigma) \geq 1) \tag{5}$$

where $N_n^i(\sigma)$ is the number of occurence of word σ in the string of length n under source model i. With respect of [ref2] result we have $K_n = C_{n,n} + O(n^{-\epsilon})$ with $C_{n,m} = \sum_\sigma (1 - (1 - p_1(\sigma))^n) \times (1 - (1 - p_2(\sigma))^m)$ where $p_i(\sigma)$ is the word probability in source model i.

We have $C_{0,m} = C_{n,0} = 0$ and for $n, m \geq 1$ we have the recursion

$$C_{n,m} = 1 + \sum_{k,\ell} \left(\begin{array}{c} n \\ k \end{array} \right) p_1^k q_1^{n-k} \left(\begin{array}{c} m \\ \ell \end{array} \right) p_2^\ell q_2^{m-\ell} (C_{k,\ell} + C_{n-k,m-\ell}) \tag{6}$$

where (p_i, q_i) is the vector of symbol probabilities in model i (p_i for symbol "0" and q_i for symbol "1", $p_i + q_i = 1$). In passing we get $C_{1,1} = \frac{1}{1 - p_1 p_2 - q_1 q_2}$.

Using the double Poissonized generating function: $C(z_1, z_2) = \sum_{n,m} C_{n,m} \frac{z_1^n z_2^m}{n!m!} e^{-z_1-z_2}$ yields

$$C(z_1, z_2) = C(p_1 z_1, p_2 z_2) + C(q_1 z_1, q_2 z_2) + (1 - e^{-z_1})(1 - e^{-z_2}) \ . \tag{7}$$

It turns that $C_{n,m} = [z_1^n][z_2^m]C(z_1, z_2)e^{z_1+z_2}$, which is a double depoissonization operation. We will prove that $C_{n,n} = C(n, n) + O(1)$ via double depoissonization.

4.1 Double depoissonization

We can prove that when z_1 and z_2 belongs to a cone $\mathcal{C}$ around the positive real axis then $C(z_1, z_2) = O(|z_1| + |z_2|)$. We show that when $|z_1|, |z_2| \to \infty$:

- if $z_1, z_2 \in \mathcal{C}$: $C(z_1, z_2) = O(|z_1| + |z_2|)$;
- if $z_1, z_2 \notin \mathcal{C}$: $C(z_1, z_2)e^{z_1+z_2} = O(e^{\alpha|z_1|+\alpha|z_2|})$ for some $\alpha < 1$;
- if $z_i \in \mathcal{C}$ and $z_j \notin \mathcal{C}$ for $\{i, j\} = \{1, 2\}$: $C(z_1, z_2)e^{z_j} = O(|z_i|e^{\alpha|z_j|})$.

Under this hypothesis one can show that with depoissonization lemmas [5] $C_n(z_2) = n![z_1^n]C(z_1, z_2)e^{z_2}$ satisfies $C(n, z_2) + g(n, z_2)$ where $g(n, z_2) = O(1 + \frac{z_2}{n})$ when $z_2 \in \mathcal{C}$ and $g(n, z_2) = O(e^{\alpha|z_2|})$ when $z_2 \notin \mathcal{C}$. Applying again depoissonization lemma on $C_n(z_2)$ and on $g(n, z_2)$, we get $C_{n,m} = C_n(m) + O(\frac{n}{m} + 1) = C(n, m) + O(\frac{n}{m} + \frac{m}{n})$.

4.2 Common words analysis under same model

We assume here that $p_1 = p_2 = p$ and $q_1 = q_2 = q$: the source models are identical. In this case we can define $c(z) = C(z, z)$ which satisfies:

$$c(z) = c(pz) + c(qz) + (1 - e^{-z})^2 \tag{8}$$

We will prove that $C_{n,n} = n\frac{2\log 2}{h} + Q(\log n)n + o(n)$, with $h = -p\log p - q\log q$ the entropy rate. Quantity function $Q()$ is a small periodic function (amplitude smaller than 10^{-6}) which is non zero only when $\frac{\log p}{\log q}$ irrational (excepted when $p = q = \frac{1}{2}$.

We quickly pass over the case $p = q = \frac{1}{2}$ (the binary uniform case) where the $c(z)$ has the simple expression $c(z) = 2z - 1 + e^{-2z}$. In this case we have $C_{n,n} = 2n + O(1)$.

We now consider the general case. Since $c(z)$ is $O(z^2)$ when $z \to 0$ and is $O(z)$ (C_n, n is smaller than the number of nodes in a trie with $2n$ independent sequences), the Mellin transform of $c(z)$, $c^*(s)$ is defined for $\Re(s) \in]-2, -1[$ and satisfies

$$c^*(s) = \frac{(2^{-s} - 2)\Gamma(s)}{1 - p^{-s} - q^{-s}} \tag{9}$$

We use the inverse Mellin $c(z) = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} c^*(s)z^{-s}ds$ for c in the definition domain of $c^*(s)$.

Using the residus theorem on the singularities of the integrand which are:

- at $s=-1$ with residus $-\frac{2z\log 2}{h}$.
- at $s=s_k$, $k\neq 0$ the denumerable set of the other roots of $1-p^{-s}-q^{-s}$ with residus $z^{-s_k}\frac{2^{-s_k}-2}{p^{-s_k}\log p+q^{-s_k}\log q}\Gamma(s_k)$.

Therefore (ignoring residus at $\Re(s)\geq 0$)

$$c(z)=\frac{2z\log 2}{h}-\sum_{k\neq 0} z^{-s_k}\frac{2^{-s_k}-2}{p^{-s_k}\log p+q^{-s_k}\log q}\Gamma(s_k)+O(1) \tag{10}$$

Notice that unless the logarithms of the probabilities are integer multiple (*i.e.* $\frac{\log p}{\log q}$ is rational in the binary case) we have $\forall k\neq 0:\ \Re(s_k)>-1$ but $\liminf \Re(s_k)=-1$. Therefore the leading term is $\frac{2\log 2}{h}n$. When $\frac{\log p}{\log q}$ is rational then there is a subsequence of the s_k that are regularily spaced on the vertical axis $\Re(s)=-1$. This impacts the leading term with a small periodic correcting term $zQ(\log z)$.

4.3 Common words between different source models

We will show in a large set of cases and conjecture in other cases that $C(z,z)=O(z^\kappa)$ where $\kappa\leq 1$ and κ is the smallest real number such there exists $s_1,s_2>0$ such that $s_1+s_2=\kappa$ and $p_1^{s_1}p_2^{s_2}+q_1^{s_1}q_2^{s_2}=1$.

We want to use of the double Mellin transform $\int\int C(z_1,z_2)z_1^{s_1-1}z_2^{s_2-1}dz_1dz_2$, integrated for $\Re(s_1)=c_1$ and $\Re(s_2)=c_2$ with (c_1,c_2) in the definition domain of $C^*(s_1,s_2)$. But this is *a priori* no such domain since we have $C(z_1,z_2)=O(z_i)$ for all cases: $z_i\to 0$ and $z_i\to\infty$ ($i\in\{1,2\}$). To remove the deadlock we define $\tilde{C}(z_1,z_2)=C(z_1,z_2)-D(z_1,z_2)$ where $D(z_1,z_2)=z_1e^{-z_1}D_1(z_2)+z_2e^{-z_2}D_2(z_1)-C_{1,1}z_1z_2e^{-z_1-z_2}$, where $D_1(z)=\frac{\partial}{\partial z_1}C(0,z)$ and $D_2(z)=\frac{\partial}{\partial z_2}C(z,0)$.

That way $\tilde{C}(z_1,z_2)=O(z_i^2)$ when $z_i\to 0$ and double Mellin transform $C^*(s_1,s_2)$ of $\tilde{C}(z_1,z_2)$ exists for $-2<\Re(s_i)<-1$.

4.3.1 Preliminary result

The Mellin transform of $D_i(z)$ exists and is equal to $D_i^*(s)=-\frac{\Gamma(s)}{1-p_ip_j^{-s}-q_iq_j^{-s}}$, since $\frac{\partial}{\partial z_i}C(z_1,z_2)=p_i\frac{\partial}{\partial z_i}C(p_1z_1,p_2z_2)+q_i\frac{\partial}{\partial z_i}C(q_1z_1,q_2z_2)+e^{-z_i}(1-e^{-z_j})$.

Therefore by playing with Mellin transform $D_i(z)=\frac{z\log z}{h_{ij}}+O(z)$ with $h_{ij}=-p_i\log p_j-q_i\log q_j$.

4.3.2 The double Mellin transform

We have the modified functional equation:

$$\begin{aligned}\tilde{C}(z_1,z_2) &= \tilde{C}(p_1z_1,p_2z_2)+\tilde{C}(q_1z_1,q_2z_2)+(1-e^{-z_1})(1-e^{-z_2})+\\ &\quad +D(p_1z_1,p_2z_2)+D(q_1z_1,q_2z_2)-D(z_1,z_2).\end{aligned}$$

Therefore

$$C^*(s_1,s_2) = \left(\frac{1}{1-R(s_1,s_2)}+\frac{s_1}{1-R(-1,s_2)}+\frac{s_2}{1-R(s_1,-1)}+\frac{s_1s_2}{1-R(-1,-1)}\right)\times \\ \times\Gamma(s_1)\Gamma(s_2)$$

with $R(s_1,s_2)=p_1^{-s}p_2^{-s_2}+q_1^{-s_1}q_2^{-s_2}$, in passing we have $C_{1,1}=\frac{1}{1-R(-1,-1)}$. Using the inverse Mellin transform we have

$$\tilde{C}(z,z)=\left(\frac{1}{2i\pi}\right)^2\int_{\Re(s_1)=c_1}\int_{\Re(s_2)=c_2}C^*(s_1,s_2)z^{-s_1-s_2}ds_1ds_2 \tag{11}$$

We will move the integration domain of s_1 to $\Re(s_1)=c_1$ and the integration domain of s_2 to $\Re(s_2)=c_2$ for $c_1,c_2=-1+\epsilon$ with $\epsilon>0$ such that $R(s_1,s_2)$ stays far away from 1. The only potential poles encountered are at $s_1=-1$ and $s_2=-1$ and are cancealed by the additional terms.

Therefore we have equation (11) still valid for $\Re(s_1)=c_1$ and $\Re(s_2)=c_2$ for some $c_1,c_2>-1$.

On the new integration domain we can focuse only on term $\int\int\frac{\Gamma(s_1)\Gamma(s_2)}{1-R(s_1,s_2)}z^{-s_1-s_2}$, the other leading to exponentially negligible terms. Indeed the term $\frac{s_1\Gamma(s_1)\Gamma(s_2)}{1-R(-1,s_2)}$ has no pole for $\Re(s_1)>-1$, and term $\frac{s_2\Gamma(s_1)\Gamma(s_2)}{1-R(s_1,-1)}$ has no pole for $\Re(s_2)>-1$. The term $\frac{s_1s_2\Gamma(s_1)\Gamma(s_2)}{1-R(-1,-1)}$ has no pole at all.

Therefore we have

$$C(z,z)=\left(\frac{1}{2i\pi}\right)^2\int_{\Re(s_1)=c_1}\int_{\Re(s_2)=c_2}\frac{\Gamma(s_1)\gamma(s_2)}{1-R(s_1,s_2)}z^{-s_1-s_2}ds_1ds_2+O(z^{-M}) \tag{12}$$

For arbitrary $M>0$.

We will look at the simplified case where $p_1=q_1=\frac{1}{2}$ and $(p_2,q_2)=(p,q)$ general. In this case $R(s_1,s_2)=2^{s_1}R(s_2)$ with $R(s)=p^{-s}+q^{-s}$ the Reynier entropy.

Now we will move the integration line of s_1 toward positive real axis direction. We meet one pole at $s_1=0$ (residu 1). The poles of function $(1-2^{s_1}R(s_2))^{-1}$ at $s_1=-\log_2 R(s_2)+\frac{2ik\pi}{\log 2}$, for k integer (residus $-\frac{1}{\log 2}$). therefore we have:

$$\tilde{C}(z,z) = -\frac{1}{2i\pi}\int\frac{\Gamma(s_2)}{1-R(s_2)}z^{-s_2}ds_2+ \\ +\sum_k\frac{1}{2i\pi\log 2}\int\Gamma\left(-L(s_2)+\frac{2ik\pi}{\log 2}\right)\Gamma(s_2)z^{L(s_2)-s_2-\frac{2ik\pi}{\log 2}}ds_2+ \\ +O(z^{-M})$$

with $L(s) = \log_2 R(s)$ and $M > 0$ being arbitrary (but fixed to have uniformty of $O(.)$). It turns out that function $-\frac{1}{1-R(s)}$ and function $\sum_k \frac{\Gamma\left(-L(s)+\frac{2ik\pi}{\log 2}\right)}{\log 2} z^{L(s)-\frac{2ik\pi}{\log 2}}$ have same poles and their residus canceal. Indeed the root of $logR(s) - 2ik\pi$ are exactly the root s_k of $1 - R(s)$ and the residus of $\Gamma\left(L(s) - \frac{2ik\pi}{\log 2}\right)$ is the same as the residu of $\frac{1}{1-R(s)}$. Therefore the integration path can be moved toward arbitrary position. Anyhow we cannot increase too much the value of c_2 because the function $z^{L(s)-s-\frac{2ik\pi}{\log 2}}$ increases when $s \to \infty$. Therefore we can take for c_2 the value of which minimizes $L(s) - s$ between $-1, 0$. Let's call κ this minimum value. We can ignore the term in $\frac{z^{-s}}{1-R(s)}$ since $L(s) > 0$ for s between $-1, 0$.

Therefore we have

$$C(z,z) = \frac{1}{2i\pi \log 2} \int_{\Re(s)=c_2} \sum_k \Gamma\left(-L(s) + \frac{2ik\pi}{\log 2}\right) \Gamma(s) z^{L(s)-s-\frac{2ik\pi}{\log 2}} ds + O(z^{\kappa-\epsilon}) \quad (13)$$

On each term there is a saddle point at $s = c_2$ therefore each of the terms $\Gamma\left(-L(s) + \frac{2ik\pi}{\log 2}\right) \Gamma(s) z^{L(s)-s-\frac{2ik\pi}{\log 2}}$ will contribute as $\frac{z^\kappa}{\sqrt{2\pi\alpha_2 \log z}} \Gamma\left(c_1 + \frac{2ik\pi}{\log 2}\right) \Gamma(c_2) z^{-\frac{2ik\pi}{\log 2}}$ with $c_1 = -c_2 - \kappa$. This is equivalent to $\frac{z^\kappa}{\sqrt{2\pi\alpha_2 \log z}} \Gamma(c_1)\Gamma(c_2)$ with fluctuating terms of period $\log 2$.

When $\frac{\log p}{\log q}$ is rationnal, then the function $R(s)$ is periodic on the imaginary axis leading to an infinity of saddle points. Therefore the main term $\frac{z^\kappa}{\sqrt{2\pi\alpha_2 \log z}} \Gamma(c_1)\Gamma(c_2)$ got also other fluctuating terms of period not necessary multiple of $\log 2$. In conclusion the fluctuating term $z^\kappa Q(\log z)$, which can be completely characterized, may not be periodic at all.

The main term can be rewriten the following way

$$C(z,z) = \frac{z^\kappa}{\sqrt{\pi \Delta R \nabla R \log z}} \Gamma(c_1)\Gamma(c_2) + z^\kappa Q(\log z) + o(z^\kappa) \quad (14)$$

with $\Delta R = \frac{\partial^2}{\partial s_1^2} R + \frac{\partial^2}{\partial s_2^2} R$ and $\nabla R = \frac{1}{2}(\frac{\partial}{\partial s_1} R + \frac{\partial}{\partial s_2} R)$ at (c_1, c_2 roots of $1 - R(s_1, s_2)$ which minimizes $s_1 + s_2$.

Classic analysis gives for the binary alphabet:

$$\kappa = \frac{\log \frac{q_2}{q_1} \log\log \frac{q_2}{q_1} + \log \frac{p_1}{p_2} \log\log \frac{p_1}{p_2} - \log \frac{q_2 p_1}{q_1 p_2} \log\log \frac{q_2 p_1}{q_1 p_2}}{\log p_1 \log q_2 - \log p_2 \log q_1} \quad (15)$$

Figure 1 displays the various value of κ. Notice that contrary to figure κ drops well to zero when p_2 or q_2 are close to zero when $p_1 = q_1 = \frac{1}{2}$, but the convergence is very slow (in $\frac{1}{\log p_2}$.

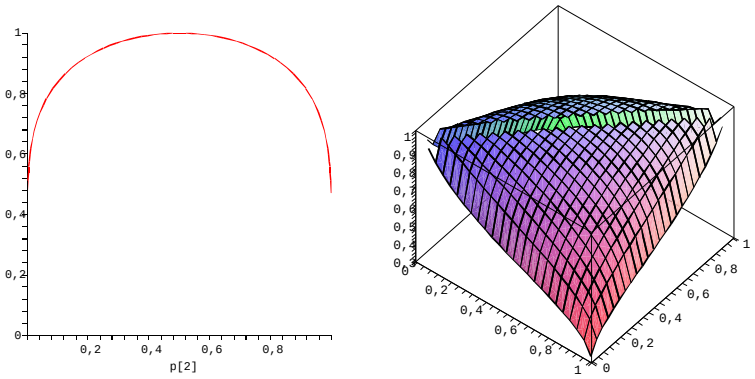

Figure 1: Quantity κ versus p_2 for $p_1 = q_1 = \frac{1}{2}$ (left) general versus (p_1, p_2) (right)

5 Number of common nodes in suffix trees

5.1 Common nodes between sequences on different models

Enumerating the common words in sequences may not as easy. An alternative method (and almost equivalent) consists into enumerating the number of common nodes in their respective suffix trees. This method has been introduced by Alberto Apostolico. Let R_n be the average number of common suffix nodes.

$$R_n = \sum_{\sigma} P(N_n^1(\sigma) \geq 2) P(N_n^2(\sigma) \geq 2) \tag{16}$$

We have $R_n = T_{n,n} + O(n^{-\epsilon})$ with $T_{n,m} = \sum_{\sigma}(1 - (1 - p_1(\sigma))^n - np_1(\sigma)(1 - p_1(\sigma))^{n-1}) \times (1 - (1 - p_2(\sigma))^m - mp_2(\sigma)(1 - p_2(\sigma))^{m-1})$ where $p_i(\sigma)$ is the word probability in source model i.

We have $T_{0,m} = T_{n,0} = T_{1,n} = T_{m,1} = 0$ and for $n, m \geq 1$ we have the recursion

$$T_{n,m} = 1 + \sum_{k,\ell} \left(\begin{array}{c} n \\ k \end{array} \right) p_1^k q_1^{n-k} \left(\begin{array}{c} m \\ \ell \end{array} \right) p_2^\ell q_2^{m-\ell} (T_{k,\ell} + T_{n-k,m-\ell}) \tag{17}$$

Using the double Poissonized generating function: $T(z_1, z_2) = \sum_{n,m} T_{n,m} \frac{z_1^n z_2^m}{n!m!} e^{-z_1 - z_2}$ yields

$$T(z_1, z_2) = T(p_1 z_1, p_2 z_2) + T(q_1 z_1, q_2 z_2) + (1 - (1 + z_1)e^{-z_1})(1 - (1 + z_2)e^{-z_2}) \ . \tag{18}$$

We have $T_{n,n} = T(n, n) + O(1)$ via double depoissonization.

The Mellin transform $T^*(s_1, s_2)$ is defined for $-2 < \Re(s_1), \Re(s_2) < -1$ and has value:

$$T^*(s_1, s_2) = \frac{\Gamma(s_1 + 1)\Gamma(s_2 + 1)}{1 - R(s_1, s_2)} \tag{19}$$

Using the same derivation as in the previous section we will get when the sequence source models differ:

$$T(z,z) = \frac{z^\kappa}{\sqrt{\pi \Delta R \nabla R \log z}} \Gamma(1+c_1)\Gamma(1+c_2) + z^\kappa Q(\log z) + o(z^\kappa) \tag{20}$$

where $Q(.)$ is a fluctuating function which arise when any of the ratio $\frac{\log p_i}{\log q_i}$ is rationnal.

5.2 Common nodes between sources on same model

Let $t(z) = T(z,z)$ when $(p_1, q_1) = (p_2, q_2) = (p, q)$, we have

$$t(z) = t(pz) + t(qz) + (1 - (1+z)e^{-z})^2 \tag{21}$$

Mellin transform of $t(z)$, $T^*(s)$ satisfies the identity

$$t^*(s) = \frac{(1+s)(\frac{1}{4}2^{-s}s - 2 + 2^{-s})\Gamma(s)}{1 - p^{-s} - q^{-s}} \tag{22}$$

and therefore $t(z) = \frac{1}{2}\frac{n}{h} + Q_5(\log n) + o(n)$ where $Q_5(.)$ is periodic when $\frac{\log p}{\log q}$ is rational.

5.3 Conclusion and perspectives

We have shown that the numbers of common words or nodes between two strings follow similar distribution. We get the unexpected results that this number is in n^κ where κ can be explicitly expessed with the parameters of the source models. In the future work we must also investigate the impact of error terms in $n^{-\epsilon}$ that arise when we neglect the impact of auto-correlation polynomials. The results have been detailled for binary alphabet but are easily extendable for any finite alphabet (although quantity κ would loose its close-formula expression). Using the standard approach used in [4], one expects to extend those results to Markov sources.

References

[1] S. Janson, S. Lonardi, W. Szpankowski, On Average Sequence Complexity, Theoretical Computer Science, 326, 213–227, 2004;

[2] P. Jacquet, W. Szpankowski, Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach, J. Combinatorial Theory. Ser. A, 66, pp. 237-269, 1994

[3] M. Régnier, W. Szpankowski, On pattern frequency occurrences in a Markovian sequence, Algorithmica , 22, 631-649, 1998.

[4] P. Jacquet, W. szpankowski Analytic Approach to Pattern Matching , Chapter 7 in Lothaire Applied Combinatorics on Words, Cambridge University Press, Cambridge, 2004. (Bibliography and Index)

[5] P. Jacquet, W. Szpankowski, Analytical depoissonization and its applications, Theoretical Computer Science, in "Fundamental Study", 201, No. 1-2, 1–62, 1998.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399